*The Canadian Journal of Program Evaluation* Vol. 22 No. 3 Pages 99–119
ISSN 0834-1516 


# ÉVALUATIONS À GRANDE ÉCHELLE DE L'ÉCRITURE : LIEN ENTRE LE SCORE HOLISTIQUE ET LES COMPOSANTES DE L'ÉCRITURE

Denis Savard
Serge Sévigny
Université Laval
Québec, Québec

Isabelle Beaudoin
Université du Québec à Rimouski
Lévis, Québec

**Résumé :** Les informations générées par les évaluations à grande échelle par l'entremise des scores holistiques en écriture devraient servir d'indicateurs aux instances décisionnelles. Mais jusqu'à quel point ces scores holistiques représentent-ils les différentes composantes de l'écriture? Se basant sur un échantillon de 3 107 productions écrites par des élèves canadiens de 13 et 16 ans, les analyses montrent que six composantes de l'écriture sont reliées au score holistique et que ces relations ne varient pas selon la langue (anglophone et francophone). Les résultats permettent de discuter des interprétations associées aux scores holistiques dans le contexte d'évaluations de l'écriture.

**Abstract:** Information generated by large-scale writing assessment holistic scores should serve as indicators for decision makers. But to what extent do holistic scores represent the different writing components? Based on a sample of written essays by 3,107 13- and 16-year-old Canadian students, analyses show that six writing components are related to the holistic score and the relationships do not vary based on language (essays written in English versus essays written in French). Results allow for a discussion of the interpretations of holistic scores in the context of writing assessments.

Depuis quelques années, plusieurs chercheurs canadiens participent à des regroupements qui permettront de cerner les problématiques associées aux évaluations à grande échelle. Cette mobilisation est issue du besoin de discuter des nombreux questionnements

Correspondance à l'auteur : Serge Sévigny, TSE 450, Département des Fondements et pratiques en éducation, Université Laval, Québec, QC, G1K 7P4; <serge.sevigny@fse.ulaval.ca>

engendrés par ce type d'évaluation et de l'orientation à donner aux recherches à venir, entre autres sur les utilisations et les interprétations qui doivent découler des résultats produits par ces évaluations.

Cette mobilisation n'est pas sans lien avec la définition du terme « validité » telle que rapportée dans la plus récente édition des *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Stipulant que la validité représente la considération la plus importante dans le développement et l'évaluation de tests, ces auteurs la définissent comme suit : « Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the uses of tests » (p. 9).

Selon Haladyna (2002), le point le plus important de tout programme d'évaluation consiste à valider l'interprétation et l'utilisation des scores des tests. Il mentionne qu'au cœur de cette validation se retrouve la documentation d'appui, résultat d'une activité coordonnée qui présente les preuves de validité à différents auditoires sous différentes formes. Toujours selon Haladyna, il y a au moins quatre raisons pour fournir de la documentation supplémentaire d'appui à la validité : renforcer la preuve de validité; obtenir des comptes rendus valides de la performance des élèves; se conformer aux standards de l'AERA, de l'APA, et du NCME; et fournir de meilleurs outils dans le cas de poursuites judiciaires. À titre d'exemple de documentation d'appui favorisant une interprétation plus juste des résultats, ou une meilleure compréhension des systèmes d'évaluation, on suggère l'addition d'études quasi-expérimentales (Tindall, 2002), la collection de mesures provenant de plusieurs sources (Hambleton, 2001; Helwig, 2002; Ryan, 2002) ou l'utilisation d'un critère externe (Kifer, 2001).

Selon Haladyna (2002), il existe une obligation des responsables de tests à grande échelle, surtout ceux impliquant les tests de réussite en éducation, à informer le public au sujet de leurs tests et de l'interprétation des scores qu'ils produisent. Il leur revient également de fournir la documentation d'appui nécessaire à la preuve de validité. Haladyna affirme que chaque interprétation de scores nécessite une validation correspondante, qu'il faut communiquer les interprétations valides et informer les gens sur celles qui ne le sont pas. Par exemple, à la suite d'une évaluation holistique à grande échelle de l'écriture, est-il juste de dire que les élèves de telle province sont meilleurs en écriture que les élèves de telle autre province?

Il semble bien que la majorité des évaluations à grande échelle ont pour but d'appuyer une variété d'interprétations dans une variété d'objectifs (Linn, 2002). Il n'est pas clair cependant que les différentes interprétations qui en découlent vont dans le même sens; ces dernières semblent plutôt confuses, voire parfois contradictoires (Gersten & Baker, 2002). Ryan et DeMark (2002) soulèvent la possibilité qu'un biais d'interprétation de la réalité puisse subsister lors des évaluations à grande échelle. Kane (1992) mentionne que la validité globale implique la création d'un argument favorisant l'interprétation choisie, argument fondé sur la base de plusieurs tests de validité. Shepard (1993) précise que, plus les enjeux d'un test sont élevés, plus les attestations de validité doivent être probantes.

Mehrens (2002) argumente à propos de l'importance d'obtenir la confirmation que les interprétations et les utilisations des tests contribuent à améliorer la réussite des élèves sans toutefois produire des résultats pervers inattendus. Par exemple, lorsque les professeurs ajustent leur enseignement afin que les élèves obtiennent un meilleur score ou que les écoles font de même pour obtenir une meilleure cote, que doit-on considérer? Doit-on conclure que les évaluations s'adaptent à l'enseignement, ou l'inverse? Voilà un exemple de situation où les tensions entre les principes théoriques idéaux et les pratiques réelles sont parfois difficiles à cerner et à éviter.

L'interprétation des résultats provenant des scores aux tests est inextricablement liée au concept de validité. Bien que les outils d'évaluation utilisés dans le cadre des programmes d'indicateurs de rendement scolaire aient fait l'objet d'études de validité afin d'assurer la bonne interprétation des résultats obtenus, peu d'études ont examiné la justesse de l'interprétation des scores holistiques obtenus par les élèves lors des évaluations à grande échelle de l'écriture. Des scores non valides peuvent entraîner des interprétations douteuses et rendre difficilement discernable le vrai du faux.

Rappelons qu'à titre d'indicateurs de rendement scolaire, les évaluations à grande échelle évaluent les compétences des élèves du secondaire en mathématique, en sciences, en lecture, et avant 2006, en écriture. Au Canada, ces évaluations couvrent le pays et leur ampleur varie selon la taille de la population étudiante et de l'échantillon sélectionné. Les échantillons dépassent fréquemment 20 000 participants selon une rotation quadriennale des matières évaluées.

Mentionnons, parmi les objectifs plus généraux des évaluations à grande échelle, l'intention d'utiliser les résultats pour établir les prio-

rités scolaires et planifier l'amélioration des programmes visant l'apprentissage des élèves. Ainsi, si on découvrait que les élèves canadiens de 16 ans récoltent un score moyen qui ne correspond pas aux attentes établies,[1] on pourrait réagir; il faudrait cependant être convaincu que ce score soit d'abord fiable, et ensuite valide quant au construit évalué, autant pour les anglophones que pour les francophones. Kifer (2001) avance qu'à défaut de servir les réformes, les évaluations à grande échelle peuvent néanmoins mesurer le niveau de réussite.

Du côté de l'évaluation de la compétence en écriture, celle-ci se mesure généralement par l'évaluation dite holistique ou globale. De nos jours, l'approche la plus populaire demeure l'évaluation holistique de l'écriture par le biais de textes écrits par les élèves (Charney, 1984; Council of Ministers of Education, Canada, 2003; Gere, 1980; Huot, 1990; Williamson & Huot, 1993). Ce type d'évaluation peut être utilisé aux différents cycles d'enseignement et d'apprentissage sous la forme de l'évaluation de l'écriture selon l'approche par compétences. Au Canada, elle consiste à établir un score variant de 1 à 5 basé sur des critères de correction dépeignant la compétence reflétée par la production écrite d'un élève. Un correcteur ou juge attribue ainsi une note selon l'impression globale laissée par le texte et non selon l'examen des composantes spécifiques de l'écriture telles que, par exemple, l'orthographe, la syntaxe, et ainsi de suite. Ce type d'évaluation fait place à des erreurs provenant des juges comme, par exemple, de mauvaises interprétations des critères de correction et des degrés différents de sévérité.

Engelhard (2002) avance que les biais des juges peuvent influencer les interprétations parce que les scores ne représentent pas une information directe mais une information dérivée d'un jugement subjectif. Dans le contexte d'évaluations à grande échelle, les scores dispensés par des juges influenceront la validité finale des interprétations (Linn, 2002). Outre la sévérité relative des juges, d'autres composantes interviennent dans la composition du score observé : la difficulté de la tâche, la difficulté du domaine, la structure de l'échelle des scores, les caractéristiques de l'élève, et les caractéristiques du système d'évaluation (Engelhard).

Certains membres de la Fédération canadienne des enseignantes et des enseignants (Canadian Teachers' Federation SAIP Working Group, 1999) ont critiqué l'utilité de l'évaluation à grande échelle de l'écriture en rapportant qu'elle ne permettait pas d'atteindre les objectifs visés par le Conseil des Ministres de l'Éducation du Canada

(CMEC). Une des critiques adressait plus particulièrement la validité du score holistique. La Fédération canadienne des enseignantes et des enseignants mentionnait que la relation entre le score holistique et les composantes de l'écriture était prise pour acquise sans même avoir été vérifiée. De ce fait, on peut se demander ce que représente vraiment un score holistique et jusqu'à quel point il est le reflet des composantes de l'écriture. Il est crucial que les recherches traitent de cette question relative à la validité de la mesure. De plus, nous ne connaissons que très peu d'études ayant fait le parallèle entre les composantes de l'écriture et le score holistique attribué à une production écrite (voir Bacha, 2001).

Cette recherche de la validité est légitime puisque les spécialistes en évaluation considèrent un instrument valide lorsqu'il mesure vraiment ce qu'il doit mesurer et lorsque les inférences qu'on peut déduire des scores obtenus sont justifiées. Est-ce vraiment le cas des inférences tirées de l'évaluation holistique? Dans le but d'éclairer le lecteur sur ce concept, il convient d'expliquer plus spécifiquement deux types de validité pertinents à l'évaluation de l'écriture : la validité de contenu et la validité de construit.

Afin d'obtenir la validité de contenu, les spécialistes examinent les items d'un test pour déterminer leur capacité à couvrir convenablement le contenu de la matière étudiée. Par exemple, un test mesurant le niveau de compétence en géographie comprend des items qui requièrent la connaissance des pays, des capitales, des régions, des climats, et ainsi de suite. En ce sens, les divers items composent le contenu du test. Or, en évaluation directe de l'écriture, le contenu du test se limite à une simple question à partir de laquelle l'élève construit sa production écrite. Ainsi, la marge d'erreur quant à la validité de contenu demeure difficile à évaluer dans une telle situation puisque l'essence même de la production écrite reflète un vaste ensemble de composantes de l'écriture, présumées comme étant maîtrisées au moment de la production écrite; bien qu'il se rapporte effectivement à l'écriture, le contenu mesuré devient immense.

Par ailleurs, la validité de construit, définie comme étant la congruence entre le construit évalué et le construit sur lequel la technique d'évaluation se base (Elliot, Plata, & Zelhart, 1990), ne fait aucun doute en évaluation holistique de l'écriture (Sireci & Gonzalez, 2003). En effet, la production écrite reflète très bien le construit associé aux habiletés de base en écriture. Quoi de mieux que d'évaluer le texte écrit par un rédacteur pour effectuer des inférences sur ses compé-

tences en écriture? Cependant, même si un test d'écriture mesure le construit visé, les scores provenant de l'évaluation holistique peuvent, quant à eux, refléter différentes caractéristiques du très vaste construit que représente l'écriture. Ainsi, la fidélité et la validité de ces scores doivent nécessairement faire l'objet d'une attention particulière en raison des inférences qui en découleront.

Bien que la fidélité de la mesure holistique, surtout si cette dernière est unique, ne fasse pas l'unanimité (voir Hayes, Hatch, & Silk, 2000), cet article aborde plutôt le thème de validité. Même si, en toute logique, les scores holistiques devraient être le reflet des composantes de l'écriture (Bacha, 2001), les recherches antérieures ne traitent guère de la validité des inférences qui découlent de l'interprétation des scores holistiques de l'écriture (voir Hayes, Hatch, & Silk). Pour qu'un score holistique soit valide, il doit d'abord bien représenter le concept d'écriture dans lequel s'inscrivent notamment les composantes de l'écriture, les connaissances du scripteur (connaissances de la langue, du texte, du monde, etc.), ses stratégies de planification, de mise en texte, de révision, de correction, et de diffusion, et les techniques qu'il utilise (calligraphie et recherche dans les outils de références tels que les dictionnaires, la grammaire, etc.)

Or, le nombre de composantes à évaluer dans l'approche globale et la difficulté pressentie à interpréter correctement les critères d'évaluation (Canadian Teachers' Federation SAIP Working Group, 1999) peuvent être révélateurs d'une piètre concordance entre le score holistique et les différentes composantes de l'écriture. Williamson (1993) abonde aussi en ce sens en mentionnant que les nombreuses interprétations qu'on peut donner aux critères de l'évaluation globale peuvent apporter une limite importante à la signification des scores.

Le but de la recherche exécutée consiste à vérifier si les composantes de l'écriture sont en relation avec le score holistique, autant chez les élèves anglophones que chez les élèves francophones du Canada (nous avons jugé pertinent de vérifier la situation dans chaque langue officielle du Canada). Si c'est le cas, le score holistique mesurera ainsi ce qu'il est réputé mesurer : la compétence en écriture. Si tel n'est pas le cas, alors les décideurs devront y aller de prudence dans l'utilisation des résultats. Un manque de corrélation signifierait à lui seul que les inférences découlant des scores holistiques ne reflètent aucunement les composantes de l'écriture et jetterait un sérieux doute quant à la validité de construit de l'évaluation holistique (ainsi, plus le score holistique est élevé, plus on s'attend à ce que les scores aux

composantes de l'écriture soient élevés). Dans le cas où le score holistique serait en faible corrélation avec les composantes de l'écriture, cela soulèverait un grave problème d'interprétation, à savoir ce que représente réellement un score holistique. Par contre, plus le score holistique serait relié à une composante de l'écriture en particulier, mieux cette composante prédirait le score holistique en question et, donc, plus elle aurait d'importance dans la variance de ce dernier.

Le problème tient donc au fait d'évaluer la signification du score holistique en le plaçant en relation avec quelques composantes de l'écriture : le contenu, la situation de communication, l'organisation, les règles de la langue, le vocabulaire, et la structure de la phrase.

Chacune des composantes (parfois appelées composantes analytiques) reflète l'atteinte d'une compétence de l'écriture. Par exemple, la composante « contenu » réfère à l'habileté à communiquer, expliciter, développer, illustrer, et intégrer les idées; la composante « situation de communication » renvoie à l'habileté à établir et maintenir un rapport avec le destinataire ainsi qu'à imprégner son texte d'un ton qui sollicitera le lecteur; la composante « organisation » se rapporte à l'habileté à agencer et ordonner le texte; la composante « règles de la langue » réfère à la maîtrise de l'orthographe et de la ponctuation; la composante « vocabulaire » concerne l'habileté à utiliser des mots et des expressions pour communiquer; et finalement la composante « structure » réfère à la maîtrise de la syntaxe et de la construction de phrase.

Ainsi, la présente étude sert d'outil de recherche d'éléments empiriques qui permettront de renseigner le lecteur sur la teneur d'un score holistique. Sans de tels renseignements, nous risquons de nous engager dans un labyrinthe d'interprétations relevant davantage de la conjecture à laquelle il est facile d'adhérer sans réticence. C'est pourquoi les questions de recherches suivantes ont été examinées : (a) Pour chaque langue, y a-t-il un lien entre le score à chaque composante de l'écriture et le score holistique? et (b) Pour chaque langue, jusqu'à quel point le score holistique est-il prédit par l'ensemble des scores aux composantes de l'écriture?

## MÉTHODE

### Participants

Au total, 3 107 productions écrites provenant du Programme d'Indicateurs de Rendement Scolaire (PIRS) de 1994 ont reçu un score ho-

listique (évaluation globale) et six scores analytiques (un score pour chacune des six composantes à l'étude). L'échantillon francophone se compose de 1 461 productions écrites et l'échantillon anglophone en comprend 1 646. Ces productions ont été sélectionnées aléatoirement parmi les 29 000 productions évaluées en 1994, et s'avèrent les seules à avoir reçu six scores analytiques et un score holistique. Les participants de 1994, 29 000 élèves de 13 et 16 ans, garçons et filles, furent sélectionnés par les administrateurs du PIRS de manière à obtenir un échantillon représentatif de la population canadienne. À cette fin, les administrateurs ont tenu compte des variations au sein des populations scolaires d'une province à l'autre. L'utilisation de plusieurs variables de stratification telles que la région géographique, le statut public ou privé, la situation urbaine ou rurale et le niveau académique des programmes ont permis d'obtenir un échantillon représentatif. Les variables de stratification variaient d'une province à l'autre selon les besoins. Pour plus de détails, le lecteur est prié de consulter le rapport du CMEC (Council of Ministers of Education, Canada, 1995, p. 27).

Procédure

La tâche administrative prévoyait plusieurs étapes séquentielles qui permettraient en bout de ligne de colliger les données recherchées. Des guides d'information et d'administration furent distribués à tous les intervenants impliqués dans le processus d'évaluation. Ces guides décrivaient les méthodes standardisées selon lesquelles l'évaluation devait se dérouler dans chaque école. Les professeurs, les parents, et les élèves ont été informés d'avance par écrit des procédures à suivre et à respecter lors de l'évaluation.

Une semaine avant l'évaluation, les élèves choisis reçurent le Recueil de textes et d'illustrations qui renfermait des articles, des anecdotes, des récits, et des bandes dessinées traitant du thème sur lequel porterait la tâche d'écriture (l'héroïsme). Ce recueil visait à stimuler le développement des idées en rapport avec le thème proposé. Il contenait également une page où l'élève pouvait inscrire ses réflexions personnelles afin de pouvoir les consulter le jour de l'évaluation, le recueil lui-même ne devant pas être apporté dans la salle de rédaction. Au verso de cette page, on trouvait, à l'intention de l'élève, une liste de points à réviser et à corriger.

Lors de l'évaluation, on accordait 2 heures 30 minutes à l'élève afin qu'il s'acquitte de la tâche d'écriture consistant à rédiger un texte de

deux à quatre pages à double interligne sur le thème de l'héroïsme. Pendant cette période, l'élève pouvait planifier son texte, l'écrire, le réviser, et le mettre au propre avant de répondre à un questionnaire complémentaire portant sur des renseignements démographiques. De plus, afin de l'aider dans sa tâche, on lui fournissait toutes les ressources matérielles habituellement disponibles en classe : dictionnaire, grammaire, guide de rédaction, recueil de conjugaisons, ordinateur, et ainsi de suite. Les élèves ne recevaient ni note, ni rétroaction pour la réalisation de ce travail.

Mesures

Les responsables de l'administration du PIRS ont formé les correctrices et les correcteurs en fonction de chacune des approches holistiques et analytiques; cette formation impliquait bien sûr un rappel, voire une mise à jour, des critères associés à l'évaluation holistique ainsi qu'à l'évaluation des six composantes de l'écriture.

Lors de cette formation, le CMEC a tout mis en œuvre pour tenter d'uniformiser la correction des productions écrites autant dans une langue que dans l'autre. Des enseignantes et des enseignants d'élèves de 13 ans et de 16 ans ont corrigé les textes des élèves. Ces enseignants représentaient les provinces et les territoires participants. Avant la session de correction, un sous-comité de ces enseignants a reçu une formation spéciale et a choisi des exemples de copies-types qui illustraient les critères de chacun des cinq niveaux établis en fonction de l'approche globale et des cinq niveaux de chacune des composantes de l'approche analytique.

La sélection des copies-types se voulait l'étape la plus importante du processus de correction puisque ces copies allaient devenir le principal outil de travail des juges et illustraient des normes dites comparables et communes aux deux groupes linguistiques pour chaque niveau de performance. Lors de l'évaluation réelle des textes, les juges devaient se poser les trois questions suivantes : (a) Quel ensemble de critères de performance décrit le mieux la qualité du texte de l'élève?; (b) À quelle copie-type ce texte ressemble-t-il le plus?; et (c) Puis-je soutenir mon jugement à la lumière des critères et des copies-types? Une fois la sélection des copies terminée et approuvée, les juges ont été entraînés à l'utilisation des critères pour placer chaque production écrite au bon niveau, d'abord pour l'approche globale puis, pour l'approche analytique (les critères étant adaptés d'une composante à l'autre).

Notons que lors de l'évaluation des composantes analytiques, les juges ne connaissaient pas le score holistique, ni les autres scores analytiques reçus par le texte. De plus, les probabilités étaient très faibles, voire quasi inexistantes, qu'un juge corrige une même copie tout en se rappelant du score qu'il lui avait décerné. De cette façon, il paraît impossible d'avancer que les juges aient pu tenter de « jumeler » les scores analytiques au score holistique.

Les juges ont été soumis à plusieurs exercices de correction supervisés afin de s'assurer de la qualité de leurs jugements. De plus, les juges devaient se plier à des règlements stricts d'évaluation visant une plus grande impartialité. Lors de l'évaluation réelle, les juges suivaient des procédures précises; essentiellement, ils devaient lire un texte faisant partie d'une pile de 30 textes puis lui accorder un score avant de passer au texte suivant et à la pile suivante (selon le rapport technique du CMEC, « les cahiers étaient groupés en lots de 30, mêlés par âge et par province ou territoires participants »). Les évaluations duraient de 10 à 15 jours consécutifs.

Suite à une vérification de la fidélité interjuges en mesurant le pourcentage d'accord auprès de 200 copies de chaque groupe linguistique, le CMEC (Council of Ministers of Education, Canada, 1995) considéra que l'attribution des niveaux holistiques de rendement était stable (fidèle) : accord de 62,6 % pour les juges anglophones et 64,6 % pour les correcteurs francophones (il s'agit d'accords exacts et non avec une marge d'erreur). Le CMEC jugea que ces données témoignaient, jusqu'à un certain point, de l'efficacité de la formation dispensée aux juges. Mentionnons que la méthode d'évaluation de l'accord interjuges choisie par le CMEC se voulait conservatrice.

Variable démographique

Mentionnons que chaque niveau de la variable *langue* (francophone et anglophone) représente des élèves qui écrivent dans leur langue maternelle (première langue apprise et encore parlée à la maison).

Variables dépendantes

*Approche globale : score holistique*. Le score global ou holistique assigné par un juge du PIRS à chaque production écrite permet de la classer à un des cinq niveaux de performance en écriture. Plus un écrit laisse une bonne impression au juge, plus cet écrit obtient un score élevé. Le Tableau 1 présente l'échelle ayant servi à l'évaluation holistique de l'écriture de 1994.

*Approche analytique : score analytique*. Les productions écrites ont reçu un score pour chacune des six composantes de l'écriture (le contenu, la situation de communication, l'organisation, les règles de la langue, le vocabulaire, et la structure de phrase). Ces six scores, tout comme ceux de l'approche globale, variaient de 1 à 5, ce dernier constituant toujours le meilleur résultat possible. L'évaluation analytique examine donc plus spécifiquement chaque composante de l'écriture et permet d'obtenir une meilleure précision sur les forces et faiblesses d'une production écrite. À titre d'exemple, le Tableau 2 présente l'échelle ayant servi à l'évaluation de la composante *règles de la langue*. Selon Bacha (2001), les scores aux composantes de l'écriture (scores analytiques) font preuve de validité de contenu, de validité de construit, et de validité concourante avec des mesures évaluant la performance en écriture. Bacha soulève que pour obtenir une rétroaction plus spécifique quant aux habiletés en écriture, l'évaluation analytique se distingue de l'évaluation holistique. Quelques auteurs soulèvent que cette dernière cible ce que l'élève fait correctement plutôt que les faiblesses qu'il éprouve (Charney, 1984; Elbow, 1999; White, 1994).

**Tableau 1**
**Niveaux de l'échelle holistique tels qu'utilisés par le CMEC en 1994**

| Niveau | Description |
|---|---|
| 1 | L'élève manie les composantes de base de l'écriture de façon rudimentaire et incertaine. L'assimilation de ces composantes n'est pas évidente. Le texte dégage une impression d'extrême simplicité, de fragmentation, ou de non fini (ou de plusieurs de ces éléments à la fois). |
| 2 | L'élève manie les composantes de l'écriture de façon incertaine et inégale. Certains des éléments sont mieux assimilés mais le développement demeure sommaire et irrégulier. Le texte dégage une impression de simplicité ou d'inégalité. |
| 3 | L'élève possède les diverses composantes de l'écriture. En général, le texte est unifié. Le développement est fonctionnel, général, et se tient généralement jusqu'à la fin. Le texte dégage une impression de clarté. |
| 4 | L'élève manie les diverses composantes de l'écriture avec efficacité. Le texte est unifié, le développement est clair et complet; l'ensemble forme un tout structuré. Le texte dégage une impression de sérieux. |
| 5 | L'élève manie les diverses composantes de l'écriture avec assurance et efficacité. Le texte est parfaitement unifié et le développement est précis et complet; les éléments se renforcent les uns les autres. Le texte dégage une impression de perspicacité et de recherche. |

**Tableau 2**
**Niveaux de l'échelle analytique tels qu'utilisés par le CMEC en 1994 pour la composante « règles de la langue » (orthographe et ponctuation)**

| Niveau | Description |
|---|---|
| 1 | Les erreurs sont fréquentes et maladroites au point de nuire à la communication. Le texte ne témoigne que d'une connaissance très partielle de certaines règles de la langue. |
| 2 | Les erreurs sont source de distractions et sont en nombre suffisant pour entraver la communication. Le texte témoigne d'une connaissance limitée ou inconsistante des règles de la langue. |
| 3 | Des erreurs mineures dans un texte peu complexe ne gênent pas la communication ou encore, plusieurs erreurs importantes dans un texte relativement compliqué ne créent pas d'obstacles majeurs à la communication. Le texte dénote une connaissance générale des règles de la langue. |
| 4 | Quelques erreurs mineures, souvent commises par inadvertance à l'intérieur d'un texte relativement compliqué, ne semblent pas entraver la communication. Le texte témoigne d'une très bonne connaissance des règles de la langue. |
| 5 | L'absence relative d'erreurs est impressionnante au regard de la complexité du texte. Le texte témoigne d'une excellente connaissance des règles de la langue. La communication est riche. |

## RÉSULTATS

### Corrélations (question 1)

Le Tableau 3 montre que chaque composante de l'écriture est en corrélation positive et élevée avec le score holistique ainsi qu'avec les autres composantes, et ce, pour les deux groupes linguistiques. Notons que la grandeur de l'échantillon influence les coefficients de corrélations. De plus, les corrélations entre le score holistique et les composantes ne varient pas selon la langue. Les analyses font également ressortir que certaines corrélations intercomposantes chez les francophones sont plus élevées que les corrélations correspondantes chez les anglophones.

### Régressions (question 2)

Eu égard au caractère ordinal des scores holistiques, une régression logistique polychotomique ordinale aurait, en principe, mieux convenu qu'une régression multiple classique. Mais compte tenu du nombre de catégories (5), de l'approximation habituellement raisonnable de la régression multiple, et surtout de la grande facilité d'interprétation

de cette dernière par opposition à la régression ordinale, le recours à l'approche classique semble justifiable afin de déterminer si les six composantes analytiques, prises ensemble, permettent de prédire le score holistique. L'équation complète de l'approche classique, en plus de fournir l'importante statistique du coefficient de détermination ($R^2$), procure également le test $t$ pour chaque coefficient (de composante) ainsi que les coefficients standardisés, le tout facilitant grandement l'interprétation de l'importance individuelle des composantes. Notons qu'aucune observation (donnée extrême) n'a eu une influence indue sur les résultats des analyses de régression multiple.

**Tableau 3**
**Matrice de corrélations de Pearson entre les composantes à l'étude et le score holistique[a]**

| | Score holistique | Situation | Contenu | Organisation | Vocabulaire | Structure | Règles |
|---|---|---|---|---|---|---|---|
| **Français ($n$ =1445)[b]** | | | | | | | |
| Score holistique | 1 | **,522** | **,504** | **,489** | **,562** | **,561** | **,581** |
| Score hol. Spearman | 1 | **,514** | **,499** | **,479** | **,560** | **,560** | **,578** |
| Situation de comm. | | 1 | ,799[c] | ,771[e] | ,709[c] | ,704[d] | ,715[e] |
| Contenu | | | 1 | ,758 | ,724[c] | ,705 | ,692 |
| Organisation | | | | 1 | ,699[d] | ,719[e] | ,687[c] |
| Vocabulaire | | | | | 1 | ,794[d] | ,781[e] |
| Structure | | | | | | 1 | ,814 |
| Règles | | | | | | | 1 |
| **Anglais ($n$ =1638)[b]** | | | | | | | |
| Score holistique | 1 | **,526** | **,537** | **,515** | **,550** | **,570** | **,559** |
| Score hol. Spearman | 1 | **,503** | **,508** | **,490** | **,527** | **,550** | **,529** |
| Situation de comm. | | 1 | ,771 | ,700 | ,671 | ,655 | ,652 |
| Contenu | | | 1 | ,758 | ,691 | ,672 | ,670 |
| Organisation | | | | 1 | ,637 | ,639 | ,651 |
| Vocabulaire | | | | | 1 | ,753 | ,728 |
| Structure | | | | | | 1 | ,795 |
| Règles | | | | | | | 1 |

[a] Toutes les corrélations sont significatives à $p < ,001$ (bilatéral).[b] Le nombre de sujets varie de 1 445 à 1 467 chez les francophones et de 1 638 à 1 647 chez les anglophones.[c d e] Indiquent qu'il y a une différence significative entre cette corrélation et la corrélation correspondante de l'autre langue;[c] $p < ,05$, [d] $p < ,01$, [e] $p < ,001$. Notez que la taille de l'échantillon peut être responsable des corrélations élevées et que les différences entre les langues, bien que statistiquement significatives, paraissent minimes au sens pratique. Pour déterminer les différences entre les langues, nous avons utilisé le logiciel offert sur le site : <http://www.medcalc.be/download.php>. Des corrélations supplémentaires ont été effectuées avec la méthode de Spearman et ces dernières présentent des coefficients similaires aux coefficients de Pearson. À titre indicatif, pour chaque langue, les coefficients de Spearman concernant de score holistique sont affichés sur la deuxième ligne.

Les analyses de régression multiple ont permis de constater que les six composantes analytiques prises ensemble permettent de prédire le score holistique, autant chez les francophones ($R^2$ = ,38) que chez les anglophones ($R^2$ = ,40; Tableau 4).

**Tableau 4**
**Coefficients de régression linéaire multiple de chaque composante chez les échantillons des deux langues**

| Langue | Modèle | *B* | *bêta* | *t* | Sig. |
|---|---|---|---|---|---|
| Anglais | Constante | ,785 | | 11,086 | ,000 |
| (*n* = 1 634) | Structure | ,170 | ,175 | 4,925 | ,000 |
| | Règles | ,130 | ,134 | 3,879 | ,000 |
| | Vocabulaire | ,146 | ,130 | 3,963 | ,000 |
| | Situation | ,101 | ,103 | 3,145 | ,002 |
| | Organisation | ,089 | ,093 | 2,947 | ,003 |
| | Contenu | ,094 | ,096 | 2,689 | ,007 |
| | | $R$ = ,634, | $R^2$ = ,402, | *adj.* $R^2$ = ,399, | $F$ = 182,062, $p$ < ,001 |
| Langue | Modèle | *B* | *bêta* | *t* | Sig. |
| Français | Constante | ,885 | | 13,641 | ,000 |
| (*n* = 1 441) | Règles | ,208 | ,250 | 6,233 | ,000 |
| | Vocabulaire | ,142 | ,155 | 3,956 | ,000 |
| | Structure | ,111 | ,126 | 3,059 | ,002 |
| | Situation | ,104 | ,111 | 2,782 | ,005 |
| | Contenu | ,031 | ,033 | 0,836 | ,403 |
| | Organisation | ,005 | ,005 | 0,134 | ,893 |
| | | $R$ = ,614, | $R^2$ = ,377, | *adj.* $R^2$ = ,374, | $F$ = 144,530, $p$ < ,001 |

*Note*. Selon des tests non paramétriques de Kolmogorov-Smirnov (K-S), les données ne suivent pas la loi normale. Les postulats de linéarité et de multicolinéarité ont été vérifiés au moyen du logiciel SPSS; du côté anglais, les index de conditionnement varient de 11,5 à 21,1 alors que du côté français, ils varient de 9,1 à 19,6. Lorsque ces index atteignent 30, les analyses sont confrontées à des problèmes de linéarité. Les données manquantes (moins de 1 % dans le pire des cas) n'ont pas été remplacées. Nous avons analysé les valeurs résiduelles avec le test de K-S. Ce dernier montre que chez les francophones, la normalité est respectée. Par contre, chez les anglophones, ce n'est pas le cas. Notez le grand nombre de participants inclus dans les analyses. De plus, à l'exception du nuage de points de la variable vocabulaire chez les anglophones, l'homogénéité des variances semble respectée. Notez également que nous avons conduit des corrélations et des régressions polychoriques, non présentées ici, afin de s'assurer que les résultats allaient dans le même sens que ceux rapportés dans ce document sur la base des coefficients de Pearson.

L'examen du Tableau 4 fait ressortir de légères différences dans l'ordre des composantes lorsqu'on compare les coefficients standardisés (auxquels nous référerons par *bêta)* des deux groupes linguistiques. Ainsi, les variables expliquant le score holistique n'ont pas les mêmes relations entre elles d'une langue à l'autre. En effet, les variables *structure* et *règles* figurent au haut de la liste (*bêta* plus élevé) chez

les anglophones alors que chez les francophones, ce sont les variables *règles* et *vocabulaire*.

On remarque également que, chez les francophones, les variables *contenu* et *organisation* présentent un *t* loin du seuil de signification de ,05. Cela signifie que leur ajout au modèle n'apporte rien de plus à l'information fournie par l'ensemble des quatre autres variables. En d'autres mots, les variables *contenu* et *organisation* n'augmentent pas significativement le coefficient de détermination multiple. Cela ne veut pas dire pour autant que ces variables ne prédisent pas le score holistique. Lorsque prises isolément, leur statistique *t* devient fortement significative ($p < ,001$). En fait, chacune des composantes analytiques explique de façon significative une partie de la variance du score holistique. Chez les francophones, les coefficients de détermination varient de ,24 à ,34 et sont tous significatifs à un niveau alpha de ,001. Chez leurs homologues, ils varient de ,26 à ,33 et sont également tous significatifs à un seuil alpha de ,001.

## DISCUSSION ET CONCLUSION

À première vue, le portrait dépeint par les analyses corrélationnelles n'indique d'aucune façon qu'un score holistique mérite qu'on lui attribue le qualitatif « non valide » lorsque comparé aux composantes évaluées de façon analytique. Les liens entre les composantes et le score holistique renforcent le fait que le degré de maîtrise croissant de certaines compétences en production écrite suggère également un degré de maîtrise croissant d'autres compétences. D'ailleurs, les résultats appuient cette affirmation en identifiant les liens fortement positifs entre les six composantes de l'écriture autant chez les francophones que chez les anglophones. Cette étude comble le vide relevé par la Fédération canadienne des enseignantes et des enseignants qui, dans son rapport (Canadian Teachers' Federation SAIP Working Group, 1999) mentionnait qu'aucune étude n'avait encore montré que plus un élève maîtrise une composante, plus il a tendance à maîtriser les autres. Ce fut effectivement le cas à l'intérieur de cette étude empirique.

Un détail vaut la peine d'être souligné : certaines corrélations intercomposantes chez les francophones sont statistiquement plus élevées que les corrélations correspondantes chez les anglophones. Quelle est la signification de tels résultats? La maîtrise des composantes serait-elle plus homogène chez les francophones que chez leurs homologues anglophones? D'autres études pourront tester cette hypothèse, de

préférence en utilisant des méthodes d'évaluation différentes de celles utilisées dans la présente étude. Il aurait certes été intéressant d'approfondir et d'interpréter ce phénomène comme tout ce qui touche d'ailleurs les différences interlangues dans cette recherche. Malheureusement, les études ayant touché le sujet ne sont pas légion. Les chercheurs en sont ainsi limités à des conjectures sans assises réelles dans la documentation.

Pour l'instant, cette étude expose le lien positif entre les scores holistiques et les principales composantes de l'écriture. Par conséquent, elle fournit de la documentation d'appui, tel que suggéré par Haladyna (2002), qui renouvelle l'interprétation pouvant être faite à partir du score holistique. Cela tend à confirmer la validité de construit du score holistique et, de surcroît, diminue les problèmes possibles d'interprétation de ce dernier. Par ricochet, cela permet de mieux circonscrire l'utilisation des résultats (AERA, APA, NCME, 1999) obtenus lors des évaluations à grande échelle de l'écriture et de mieux interpréter les comptes rendus de la performance des élèves. Cela peut se révéler utile aux responsables des tests et aux gens qui ont à cœur la diminution de la confusion entourant la signification d'un score holistique (Gersten & Baker, 2002). Lorsque les enjeux des tests sont élevés ou qu'ils entraînent des comparaisons entre différentes populations, il s'avère important de pouvoir compter sur des preuves de validité ou sur de la documentation d'appui (Haladyna). Les résultats de la présente étude semblent corroborer ce que Kifer (2001) avançait, à savoir que les évaluations à grande échelle de l'écriture peuvent mesurer le niveau de réussite en écriture à l'aide du score holistique; cela bien entendu, dans la mesure où le score obtenu est fidèle. Il est donc possible que les critères des grilles de correction utilisées par les juges, bien que controversés dû à un prétendu manque de clarté (Canadian Teachers' Federation SAIP Working Group, 1999), aient été suffisamment bien compris par ces derniers. En effet, le score holistique reflète les six composantes de l'écriture mentionnées dans ces critères. D'autres recherches devront s'attarder à vérifier l'exactitude de telles spéculations.

Par ailleurs, les analyses de régression multiple révèlent des coefficients de détermination fort acceptables dans les deux modèles incluant les six composantes analytiques chez les francophones et les anglophones respectivement. Faisons remarquer que ces coefficients de détermination ne peuvent pas, en théorie, avoir été biaisés à la hausse, car le juge ayant attribué le score holistique à une production donnée n'était fort probablement pas celui qui a attribué tous les scores analytiques.

En considérant un écrit dans une perspective holistique ou globale, les correcteurs ne jugent pas séparément les facteurs singuliers qui composent la pièce écrite (traitement du thème, sélection des méthodes rhétoriques, choix de mots, grammaire, et mécanique). On leur demande plutôt de considérer ces facteurs comme étant des éléments faisant équipe afin de produire une impression globale sur le lecteur (Elliot, Plata, & Zelhart, 1990). C'est cette impression globale qu'un score holistique est censé refléter. Malgré cela, il apparaît maintenant plus clairement que le score global ou holistique représente, en partie, les composantes de l'écriture. Comme le mentionnent Elliot, Plata et Zelhart, l'évaluation holistique d'un texte écrit se compare à une tentative de voir une production écrite comme étant plus que la simple somme de ses caractéristiques élémentaires.

La majorité des organisations et des spécialistes reconnaissent d'emblée l'évaluation holistique d'un texte écrit comme étant potentiellement valide (Williamson, 1993; Yancey, 1999). Cependant, tout comme White (1994) le mentionne, il importe de s'interroger sur la valeur des scores provenant d'une seule production écrite sur un thème imposé et produite dans un laps de temps parfois trop court et dans un contexte où la motivation à écrire peut faire défaut. Il convient donc que les responsables de telles évaluations usent de prudence et fassent preuve de responsabilité lors de l'interprétation des résultats liés à ce genre d'évaluation.

De plus, cette étude souffre d'une limite quant à la fidélité des mesures évaluées. Rapportons que la fidélité interjuges des évaluations de l'écriture à grande échelle du CMEC tourne autour de 0,60. Cela paraît acceptable dans de telles situations mais montre tout de même que la fidélité des scores est loin d'être parfaite. De ce fait, l'interprétation de nos résultats doit tenir compte de cette limite. La validité de construit rapportée ici n'entérine en rien la fidélité de la mesure. De ce fait, il apparaît difficile de porter un jugement totalement éclairé puisque nous ne disposons que d'une partie des évidences nécessaires à une interprétation plus solide des résultats. Voilà pourquoi d'autres études s'avèrent essentielles.

Par ailleurs, les données ont été amassées selon un ordre hiérarchique, les élèves ont été sélectionnés dans une école et les écoles, dans une province. Bien qu'une telle structure puisse fausser les résultats si elle n'est pas prise en compte, notons tout de même que nous examinons les scores obtenus par les mêmes élèves, que l'échantillon est assez volumineux, et qu'il s'avère possible que les scores soient représentatifs de ceux de la population.

Malgré cela, les résultats de cette étude augmentent les connaissances quant à la validité des interprétations faites lors d'évaluations impliquant de scores holistiques en écriture. Il apparaît important de documenter la validité des évaluations à grande échelle tant au niveau de leur notation, de l'utilisation des scores obtenus, et de leur pertinence sur le plan conceptuel. Mais ce n'est là qu'une pièce du casse-tête; l'évaluation holistique ne se limite pas à l'évaluation à grande échelle, et les résultats rapportés ici permettront de mieux juger des interprétations à donner aux évaluations holistiques de tous genres, passées et à venir. Les résultats présentés dans cet article ouvrent vers de nouvelles recherches. Celles-ci, à l'instar de la nôtre, permettront de mettre à jour des faits supplémentaires qui, de fil en aiguille, s'ajouteront les uns aux autres dans le but de cerner plus en profondeur le problème relevé ici, soit l'examen de la validité des scores holistiques obtenus par les élèves tant au Canada qu'ailleurs dans le monde.

## NOTE

1. Le CMEC tenait une session pour chaque évaluation à grande échelle de l'écriture où environ une centaine de personnes impliquées en éducation étaient invitées à déterminer quel niveau holistique les élèves de 16 ans devraient atteindre. Si les gens décidaient que les élèves de 16 ans devraient atteindre le niveau 3 sur 5, alors le niveau 3 correspondait aux *attentes* fixées par le CMEC. Il en allait de même pour les élèves de 13 ans.

## RÉFÉRENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC : American Educational Research Association

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29,* 371–383.

Canadian Teachers' Federation SAIP Working Group. (1999, novembre). *Report on the SAIP 1998 reading and writing II assessment: Implications for educational policy and practice*. Toronto, ON : Auteur.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65–81.

Council of Ministers of Education, Canada (CMEC). (1995). *School Achievement Indicators Program (SAIP 1994): Report on reading and writing assessment.* Toronto, ON: Auteur.

Council of Ministers of Education, Canada (CMEC). (2003). *School Achievement Indicators Program (SAIP 2002): Report on reading and writing assessment.* Toronto, ON : Auteur.

Elbow, P. (1999). Ranking, evaluating and liking: Sorting out three forms of judgments. Dans R. Straub (Éd.), *A sourcebook for responding to student writing* (pp.175–196). Creskill, NJ : Hampton.

Elliot, N., Plata, M., & Zelhart, P. (1990). *Program development handbook for the holistic assessment of writing*. Lanham, MD : University Press of America.

Engelhard, G. Jr. (2002). Monitoring raters in performance assessments. Dans G. Tindal & T. Haladyna (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ : Erlbaum.

Gere, A.R. (1980). Written composition: Toward a theory of evaluation. *College English, 42,* 44–48.

Gersten, R., & Baker, S. (2002). The relevance of Messick's four faces for understanding the validity of high-stakes assessments. Dans G. Tindall & T.M. Haladyna (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 49–66). Mahwah, NJ : Erlbaum.

Haladyna, T.M. (2002). Supporting documentation: Assuring more valid test score interpretations and uses. Dans G. Tindall & T.M. Haladyna, (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 89–108). Mahwah, NJ : Erlbaum.

Hambleton, R.K. (2001). The next generation of the ITC Test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164–172.

Hayes, J.R., Hatch, J.A., & Silk, C.M. (2000). Does holistic assessment predict writing performance? *Written Communication, 17*(1), pp. 3–26.

Helwig, R. (2002). A methodology for creating an alternative assessment system using modified measure. Dans G. Tindall & T.M. Haladyna

(Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 427–452). Mahwah, NJ : Erlbaum.

Huot, B. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201–213.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy*. Dans la série de T.R. Guskey & R.J. Marzano (Éds.), Experts in Assessment. Thousand Oaks, CA : Corwin.

Linn, R.L. (2002). Validation of the uses and interpretations of results of state assessment and accountability systems. Dans G. Tindall & T.M. Haladyna (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 27–48). Mahwah, NJ : Erlbaum.

Mehrens, W.A. (2002). Consequences of assessment: What is the evidence? Dans G. Tindall & T.M. Haladyna (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 149–177). Mahwah, NJ : Erlbaum.

Ryan, J.M. (2002). Issues, strategies, and procedures for applying standards when multiple measures are employed. Dans G. Tindall & T.M. Haladyna (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 289–316). Mahwah, NJ : Erlbaum.

Ryan, J.M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. Dans G. Tindall & T.M. Haladyna (Éds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 67–88). Mahwah, NJ : Erlbaum.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Sireci, S.G., & Gonzalez, E.J. (2003, avril). *Evaluating the structural equivalence of tests used in international comparisons of educational*

*achievement*. Présenté à la réunion annuelle du National Council on Measurement in Education, Chicago, IL.

Tindall, G. (2002). Large-scale assessment for all students: Issues and options. Dans G. Tindall & T.M. Haladyna (Éds), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 1–24). Mahwah, NJ : Erlbaum.

White, E.M. (1994). *Teaching and assessing writing* (2e éd.). San Francisco, CA : Jossey-Bass.

Williamson, M. (1993). An introduction to holistic scoring: The social, historical and theoretical context for writing assessment. Dans M. Williamson & B. Huot (Éds.), *Validating holistic scoring for writing assessment: Theoretical & empirical foundations* (pp. 1–43). Creskill, NJ : Hampton.

Williamson, M., & Huot, B. (1993). *Validating holistic scoring for writing assessment: Theoretical & empirical foundations*. Creskill, NJ : Hampton.

Yancey, K.B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*(3), 483–503.

**Denis Savard**, Ph. D., enseigne à la Faculté des Sciences de l'éducation de l'Université Laval. Son enseignement et ses recherches traitent de l'évaluation des programmes et de la mesure de la performance des établissements et des systèmes éducatifs.

**Serge Sévigny**, professeur au département des Fondements et pratiques en éducation de l'Université Laval, étudie la validité des comparaisons entre les scores obtenus par les anglophones et les francophones lors de lévaluation à grande échelle de l'écriture.

**Isabelle Beaudoin**, docteure en psychopédagogie, enseigne au département des sciences de l'éducation de l'Université du Québec à Rimouski. Elle poursuit des recherches en didactique du français et s'intéresse plus particulièrement à la prévention des difficultés d'apprentissage en lecture et écriture.